HAIR: Hierarchical Visual-Semantic Relational Reasoning for Video Question Answering

Supplementary Material

Fei Liu\textsuperscript{1,2} Jing Liu\textsuperscript{1,2} Weining Wang\textsuperscript{1} Hanqing Lu\textsuperscript{1,2}
\textsuperscript{1}National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{2}School of Artificial Intelligence, University of Chinese Academy of Sciences
liufei2017@ia.ac.cn \{jliu, weining.wang, luhq\}@nlpr.ia.ac.cn

A. Statistics of Datasets

In Table 7, we summarize the statistics of four VideoQA datasets used in our experiments. TGIF-QA dataset includes four tasks: Action, Transition (Trans.), FrameQA, and Count. Action and Trans. are of multiple-choice task with five answer choices per question, and FrameQA and Count are of open-ended task. MSVD-QA and MSRVT-T-QA datasets only contain open-ended questions with a pre-defined answer set of size 1,000. Questions are simply divided into five types (i.e. what, who, how, when and where) according to the initial word. Youtube2Text-QA dataset consists of the videos from MSVD video set and the question-answer pairs collected from Youtube2Text video description corpus. Both open-ended and multiple-choice tasks exist.

Table 7. Statistics of the VideoQA datasets used in our experiments. #MC: the number of answer choices for multiple-choice questions.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Video</th>
<th>#Question</th>
<th>Vocab. size</th>
<th>Ans. size</th>
<th>#MC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Val</td>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGIF-QA</td>
<td>71,741</td>
<td>125,473</td>
<td>13,941</td>
<td>25,751</td>
<td>8,000 1,746 5</td>
</tr>
<tr>
<td>MSVD-QA</td>
<td>1,970</td>
<td>30,933</td>
<td>6,415</td>
<td>13,157</td>
<td>4,000 1,000 NA</td>
</tr>
<tr>
<td>MSRVTT-QA</td>
<td>10,000</td>
<td>158,581</td>
<td>12,278</td>
<td>72,821</td>
<td>8,000 1,000 NA</td>
</tr>
<tr>
<td>Youtube2Text-QA</td>
<td>1,970</td>
<td>88,350</td>
<td>6,489</td>
<td>4,590</td>
<td>6,500 1,000 4</td>
</tr>
</tbody>
</table>

B. Visualization Results

We show more visualization results including some failure examples in Figure 8. It is seen from successful examples (shown on the left) that our model is able to focus on crucial objects and frames, and capture appropriate object-level and frame-level relationships, thus giving the correct answer. On the right we show some failure examples. The analysis is given as below.

In the first example, the answer-relevant object arm is hard to detect compared with the conspicuous object hand (or finger). Besides, when the man bounces arms, the hands are also bounced jointly. Therefore, the model tends to predict the answer related to hand (or finger). In the second example, since a regular explosion is usually followed by many scattered and asynchronous explosions, it is particularly difficult for models and even humans to precisely count the number of fireworks explosion. For the third example, although successfully retrieving the relevant objects and frames, the model mistakenly recognizes the key object mower as car. We find that mower is a rare answer in the dataset, and is similar to car in appearance. Therefore, the model may require more samples that have “mower” as the ground-truth answer to learn the fine-grained distinctions between mower and car. In the last example, the given video frames are rotated 90 degrees. This may result in that the action of “fall” is mistaken for the inverse action (e.g. “lift”), thus affecting the decision of the model. We believe these analysis of failure examples may inspire the future research on VideoQA.
C. Number of Detected Objects

We investigate the impact of different number of detected objects in Table 8. $N = 6$ attains the best performance on most tasks. Thus, by default, we use 6 detected objects per frame in experiments. $N = 2$ achieves the worst performance because some important objects may be missed. Using too many objects (e.g. $N = 10$) produces slight performance drop on most tasks. This may be due to that most samples in the dataset only rely on several salient objects to answer the question and feeding too many objects into network would confuse the model.

<table>
<thead>
<tr>
<th>#objects per frame</th>
<th>Action</th>
<th>Trans.</th>
<th>FrameQA</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 2$</td>
<td>76.0</td>
<td>81.5</td>
<td>58.6</td>
<td>4.08</td>
</tr>
<tr>
<td>$N = 6$</td>
<td><strong>77.8</strong></td>
<td><strong>82.3</strong></td>
<td>60.2</td>
<td><strong>3.88</strong></td>
</tr>
<tr>
<td>$N = 10$</td>
<td>77.2</td>
<td>82.1</td>
<td><strong>60.9</strong></td>
<td>3.93</td>
</tr>
</tbody>
</table>

D. Multi-Scale Node Aggregation

For the VideoQA task, answering different questions usually needs temporal information of different durations. For example, answering the question “How many times do the fireworks explode?” may need information across the whole video, while answering the question “How many boys are doing the wheelbarrow race and one rolls?” may only need information from a few key frames. To this end, we develop a multi-scale node aggregation method to model the temporal information of different durations. The module consists of $H$ heads, in which the core component is a 1D temporal average pooling layer with specific kernel size. We experiment with different $H$. As shown in Table 9, incorporating the multi-scale temporal information significantly improves the performance on the tasks that require strong temporal reasoning, such as
Action and Trans. tasks. It is not surprising that slight performance improvement takes place on the FrameQA task, where a single frame is sufficient to infer the answer.

<table>
<thead>
<tr>
<th>#heads (kernel sizes of temporal pooling)</th>
<th>Action</th>
<th>Trans.</th>
<th>FrameQA</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H = 1$ (1)</td>
<td>75.3</td>
<td>80.6</td>
<td>59.5</td>
<td>4.08</td>
</tr>
<tr>
<td>$H = 2$ (1, 2)</td>
<td>76.8</td>
<td>81.7</td>
<td>60.0</td>
<td>3.97</td>
</tr>
<tr>
<td>$H = 4$ (1, 2, 3, 4)</td>
<td>77.8</td>
<td>82.3</td>
<td>60.2</td>
<td>3.88</td>
</tr>
</tbody>
</table>

E. Prediction Examples on MSRVTT-QA

In Figure 9, we show some prediction examples on the MSRVTT-QA dataset that contains longer video sequences. On these examples, our model gives the correct answer.

- **Q:** How many men discussed the supreme court?  
  **A:** (Baseline) Four  
  **Ours** Two  
  **GT** Two

- **Q:** What is someone driving a car on?  
  **A:** (Baseline) Ground  
  **Ours** Road  
  **GT** Road

- **Q:** What is a man with messy hair doing?  
  **A:** (Baseline) Play  
  **Ours** Cook  
  **GT** Cook

- **Q:** What is there is a young man doing?  
  **A:** (Baseline) Visit  
  **Ours** Hike  
  **GT** Hike

- **Q:** What does a guy in a gym pick up while people dance behind him?  
  **A:** (Baseline) Box  
  **Ours** Board  
  **GT** Board

Figure 9. Prediction results on MSRVTT-QA. GT: Ground-Truth.